This appendix presents additional materials and results. First, we detail our prompts in experiments in Sec. A to enhance comprehension. Then, we present additional experiments in Sec. B. Finally, qualitative results are presented in Sec. C.

A. Prompts Details in CC-VQA

Prompt for Parametric Context Generation:

> Here is the question: **{Question}** Please describe the image about the question by your own Knowledge.

Prompt for Disambiguated Question:

> Please refer to the given image and rewrite the question so that entities and attributes are explicit, pronouns are disambiguated, and the wording is more specific. Keep the original intent and scope unchanged; do not add new information. Only output the rewritten question wrapped in <question></question>. Original question: **{Question}**

Prompt for Visual Rationale Extraction:

> Here is the question: **{Question}**, Here is the selected section:**{section}**, Give your answer and put the feature or reason for the asnwer related to the image in <reason> </reason>.

Prompt for Visual-Centric Conflict Analysis:

> Here is the question: **{Question}**. Below are the reasons supporting the answer derived from the retrieved information: **{Reasons text}**. Identify the key features that distinguish the aforementioned categories and extract the features that need attention in <reason> </reason> without detailed illustration or answer.

Prompt for Correlation-Aware Position Encoding:

> Here is the question: **{Question}**. Here is the feature to focus on: <reason>{Features} </reason>. Here is the retrieved information: **{Retrieved Information}**. Short Answer:

B. Additional Experiments

B.1. Ablation Study of Compression Ratio τ

In this section, we present an ablation study on the choice of τ in our Correlation-Aware Positional Encoding method. Specifically, τ denotes the percentage of sentences with the lowest correlation scores selected for compression. We conducted comparative experiments on a subset of 10K samples from InfoSeek. The results demonstrate that the accuracy progressively increases as more low-correlation sentences are compressed. Based on **Observation 2** in the paper, we set $\tau = 75\%$, which compresses the bottom 75% of sentences ranked by correlation score during position encoding. This indicates that compressing low-correlation sentences enables the model to better focus on and utilize high-correlation sentences that are most likely to contain answers.

Table 1. **Ablation of Compression Ratio τ.**

Method	τ	Accuracy
Qwen2.5-VL-7B	25%	43.5
	50%	44.7
	75%	45.0

B.2. Generalization of CC-VQA

We validated the effectiveness and generalizability of our method by testing it on large-scale KB with 100M entries using Qwen3-VL-8B. The results demonstrate that our method still achieves a significant improvement of 3.1% (47.7% $\rightarrow$ 50.8%) on a stronger model. Additionally, we evaluated the impact of retrieval size on the results. Compared to the default top-3, the method can achieve an additional 1% improvement when using top-5 retrieval.

Table 2. **Generalization across Models and Retrieval Size.**

Method	Model	Acc.	Method	Retrieval Size	Acc.
Vanilla RAG	Qwen3-VL-8B	47.7	CC-VQA	top-3	50.8
CC-VQA	Qwen3-VL-8B	50.8	CC-VQA	top-5	51.8

B.3. Comparison with Thinking Model

Although CC-VQA involves multiple calls to the VLM model, each module uses the **same** model. To address concerns about test-time scaling, we compared it with different models. As shown in the table, our method achieved the highest accuracy (50.8%). Compared to using a stronger model (Qwen3-VL-8B-Thinking), our approach consumes fewer output tokens (192 vs. 817) and has lower latency (8.94s vs. 11.79s), but achieves higher accuracy. This confirms that our performance improvement stems from the proposed conflict mitigation methods.

Table 3. **Comparison with Thinking Model.**

Method	Model	Out Tok.	Latency (s)	Acc. (%)
Vanilla RAG	Qwen3-VL-8B-Instruct	75.74	5.2	47.7
Vanilla RAG	Qwen3-VL-8B-Thinking	817.93	11.79	48.8
CC-VQA	Qwen3-VL-8B-Instruct	192.90	8.94	50.8

B.4. Inference Time

In this section, we evaluate the inference time of our method. Specifically, we select a subset of 10K samples from InfoSeek for this evaluation. As shown in Table 4, compared to the CoCoA method, our approach benefits from token compression, resulting in lower per-sample inference time. Here, "s/k" denotes the time in seconds required to process 1K samples.

Table 4. **Inference Time.**

Method	Model	Inference Time (s/k)
CoCoA	Qwen2.5-VL-7B	713.3
CC-VQA **(Ours)**	Qwen2.5-VL-7B	616.4

Compared to Wiki-PRF, our method achieves comparable latency (8.94s vs. 8.77s) with only one additional forward pass (6 vs. 5), while remaining entirely training-free. With 76 GB of A800 GPU memory usage, our CC-VQA achieves an accuracy of 45.1% on InfoSeek, demonstrating its powerful capabilities.

Table 5. **Comparison with Wiki-PRF.**

Method	**Total Time**	**Analysis**	**Generation**	**Gen. Finetuning**	**Acc.(%)**
Wiki-PRF (SOTA)	8.77s	3.30s	5.34s	✔	42.8
CC-VQA (Ours)	8.94s	4.99s	3.95s	✘	45.1

C. Qualitative Results

C.1. Illustration of VCCR

The key idea of VCCR is to decompose the analysis of the model and retrieval contexts into multiple fine-grained subtasks (e.g., extraction and comparison), thereby constructing visual conflict reasoning to enhance problem-solving accuracy. To validate its effectiveness, we conducted both manual and MLLM-based verification. The result demonstrates that decomposed subtasks maintain high accuracy, while the VCCR module ultimately achieves overall accuracy of over 84%.

Table 6. **Analysis of VCCR.**

Verifier	**#Samples**	**Generation**	**Extraction**	**Analysis**	**VCCR**
Human	100	96%	77%	98%	91%
MLLM-based	10,000	93.87%	82.31%	95.55%	84.79%

Figure 1 illustrates the score distribution across 10,000 evaluated samples. Our analysis indicates that the model's evaluation criteria are comparatively more stringent than those employed by human annotators. Consequently, scores of 3 or higher can be reasonably interpreted as indicative of accurate assessments. This empirical distribution further substantiates the validity and methodological soundness of the proposed VCCR framework.

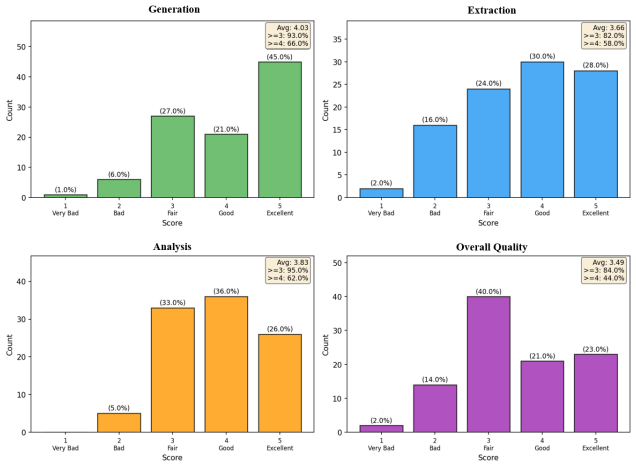

Figure 1. **Distribution of MLLM-based Verification for VCCR.**

Two sample-level case study of VCCR is presented in Figure 2. First, VCCR extracts parametric knowledge from the model based on visual inputs. It then executes reasoning on the retrieved data to isolate key features. In the final stage, all extracted features are consolidated to derive discriminative attributes, which serve as the basis for the ultimate reasoning process. The results demonstrate VCCR's ability to accurately capture fine-grained visual concepts and reasoning logic.

C.2. More Comparison Cases

In Figure 3, we present additional qualitative results of our method on Encyclopedic-VQA (E-VQA). Unlike the examples in the main text, these cases represent scenarios where both the base model and the standard RAG approach fail to produce correct answers. This further validates the effectiveness of sentence-level similarity for filtering relevant and accurate information. We specifically select three categories: animals, plants, and buildings. For each category, we show two types of questions: one indirectly related and one directly related to the image content.

In Figure 4, we present additional qualitative results of our method on Infoseek. These cases further enrich the scope of our evaluation by including examples such as the author of a jigsaw puzzle, the launch site of a rocket, and the time zone of a small town. The results demonstrate that our correlation computation effectively enhances the model's ability to extract correct answers from retrieved information.

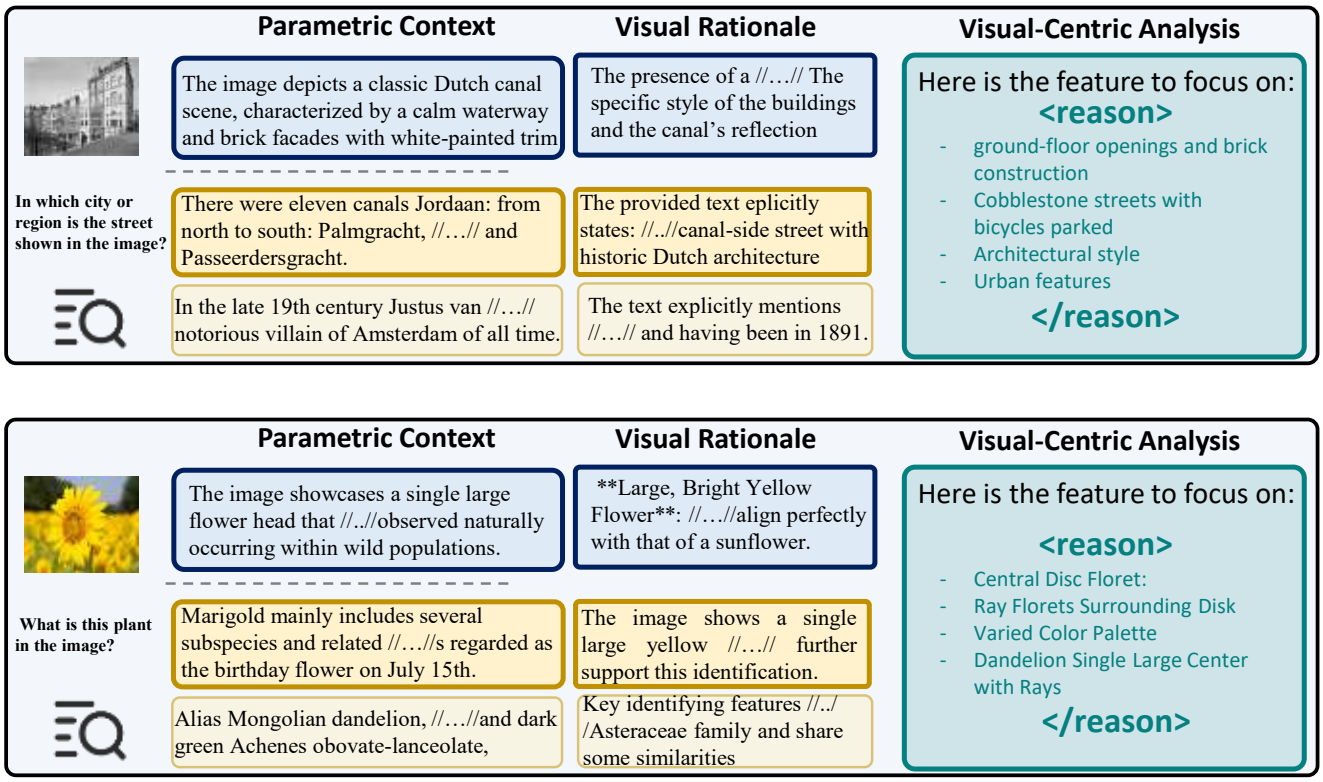

Figure 2. Detail Case of VCCR.

C.3. Illustration of CC-VQA

In this section, we mainly show the case examples of CC-VQA. Figure 5 and Figure 6 illustrate complete inference pipelines on CC-VQA. Specifically, Figure 5 presents a location-related case from E-VQA, while Figure 6 shows a time-related case from InfoSeek.

D. Limitations and Future Work.

One limitation of our method is its requirement for explicit externalization of model knowledge prior to performing visual-centric conflict reasoning. Ideally, after receiving external knowledge contexts, the model should implicitly identify and resolve conflicts between internal and external knowledge, necessitating robust reasoning capabilities. In future work, we will explore integrating multimodal reasoning into the KB-VQA task to further enhance response accuracy.

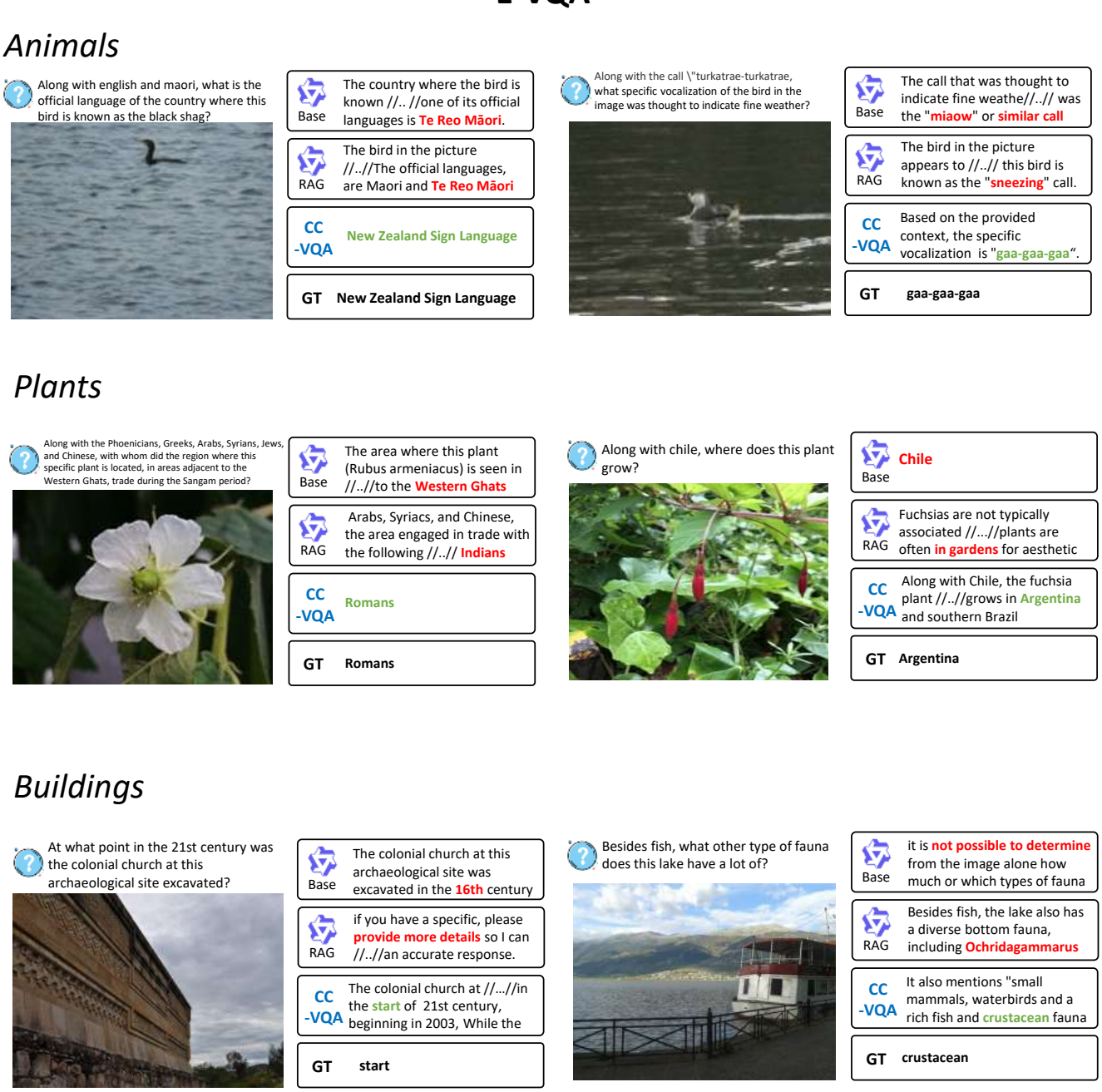

Figure 3. **More Cases of CC-VQA on E-VQA**. Qualitative results from Encyclopedic-VQA demonstrate that our method also enhances the model's retrieval-augmented generation capabilities.

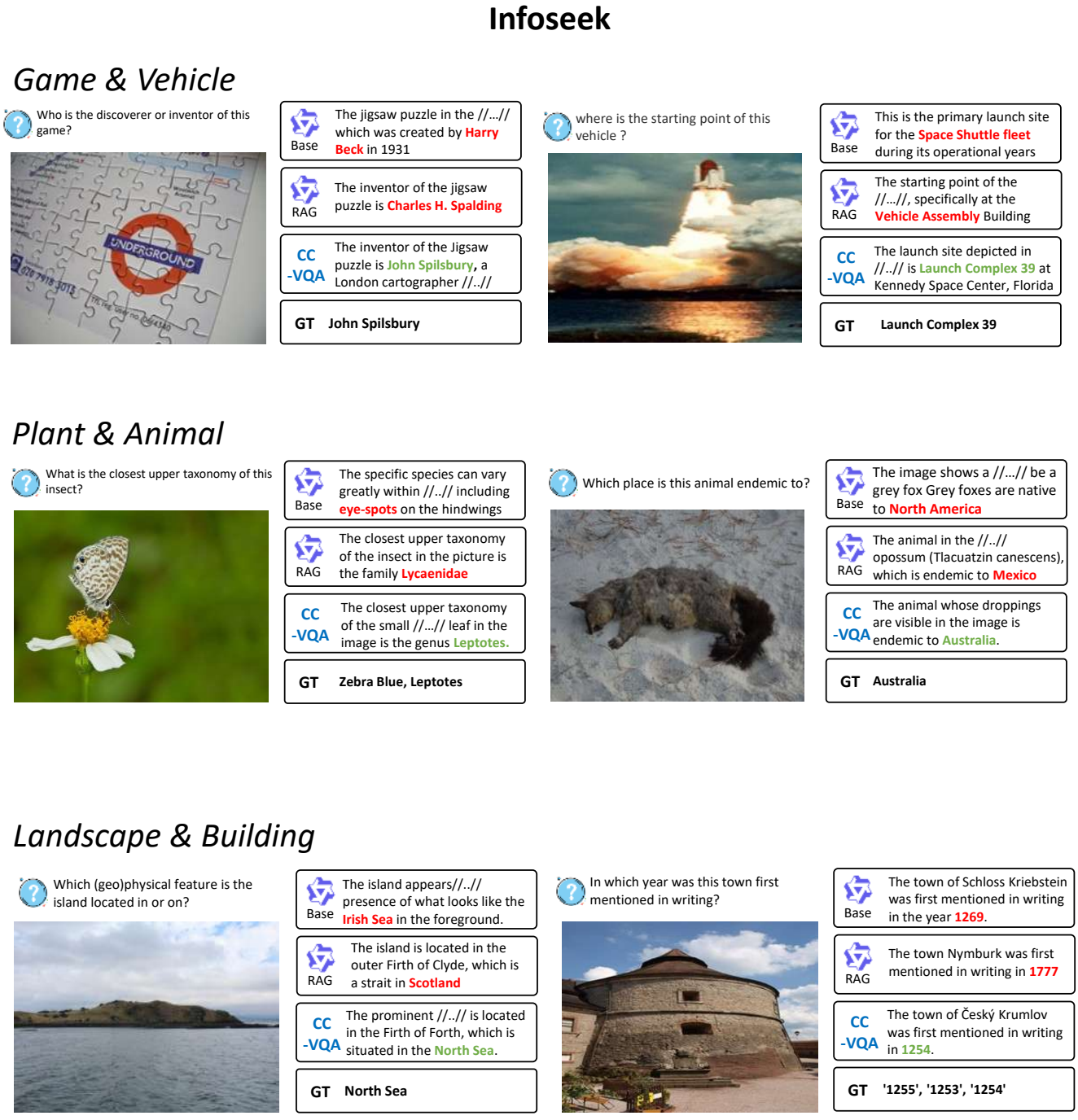

Figure 4. **More Cases of CC-VQA on Infoseek**. Qualitative results from Infoseek demonstrate that our method also enhances the model's retrieval-augmented generation capabilities.

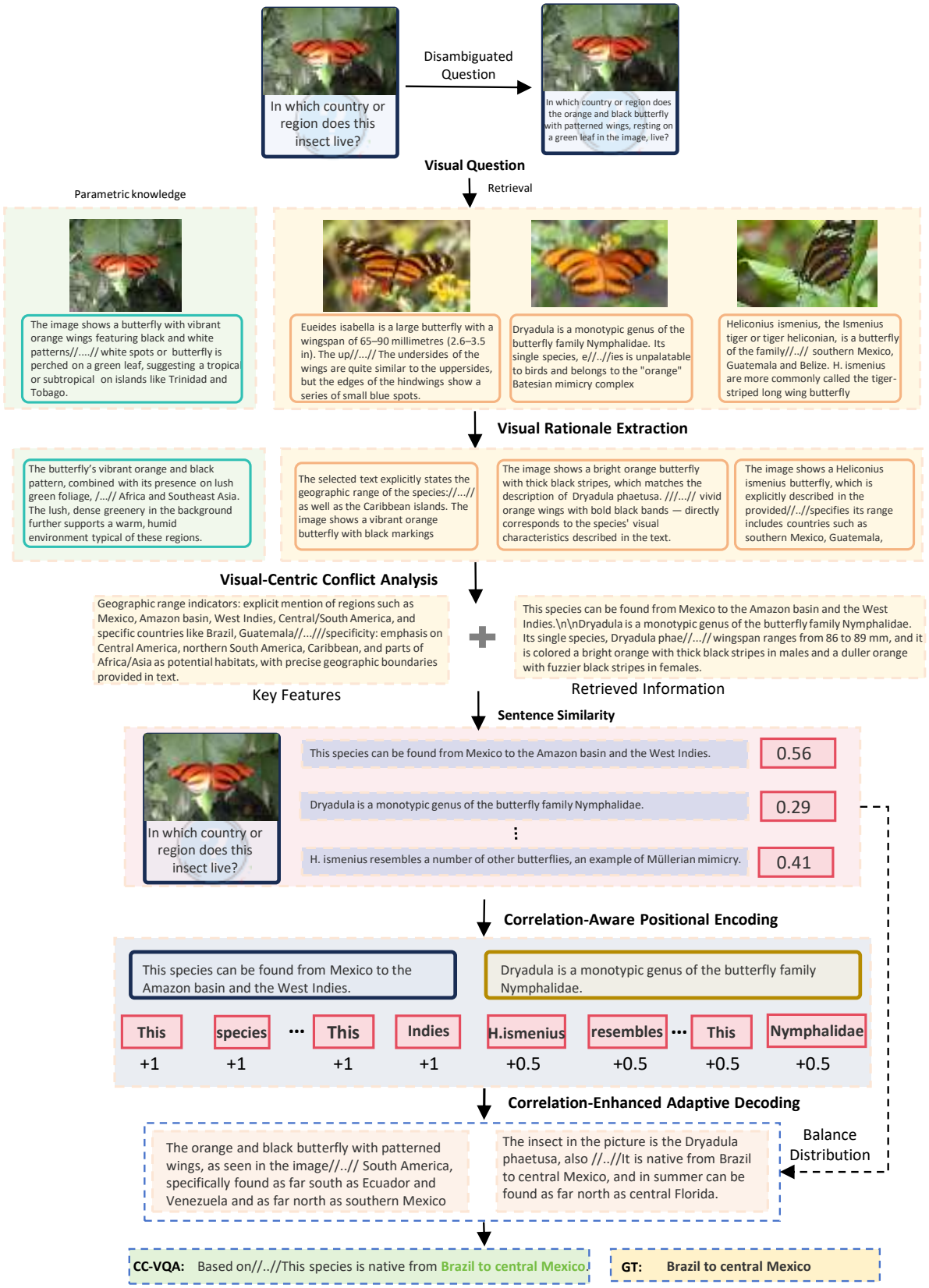

Figure 5. **Illustration of CC-VQA on E-VQA.**

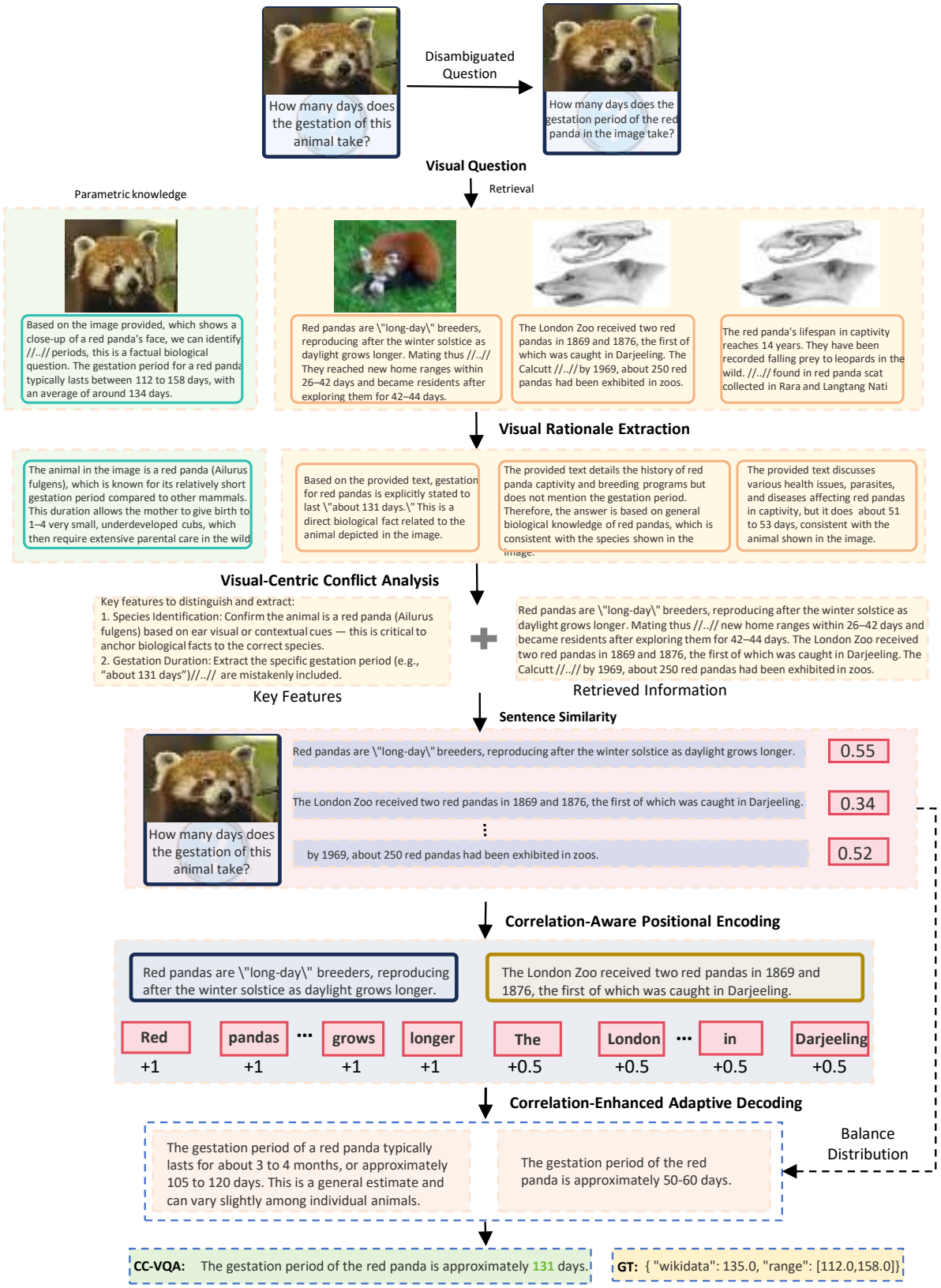

Figure 6. **Illustration of CC-VQA on Infoseek.**